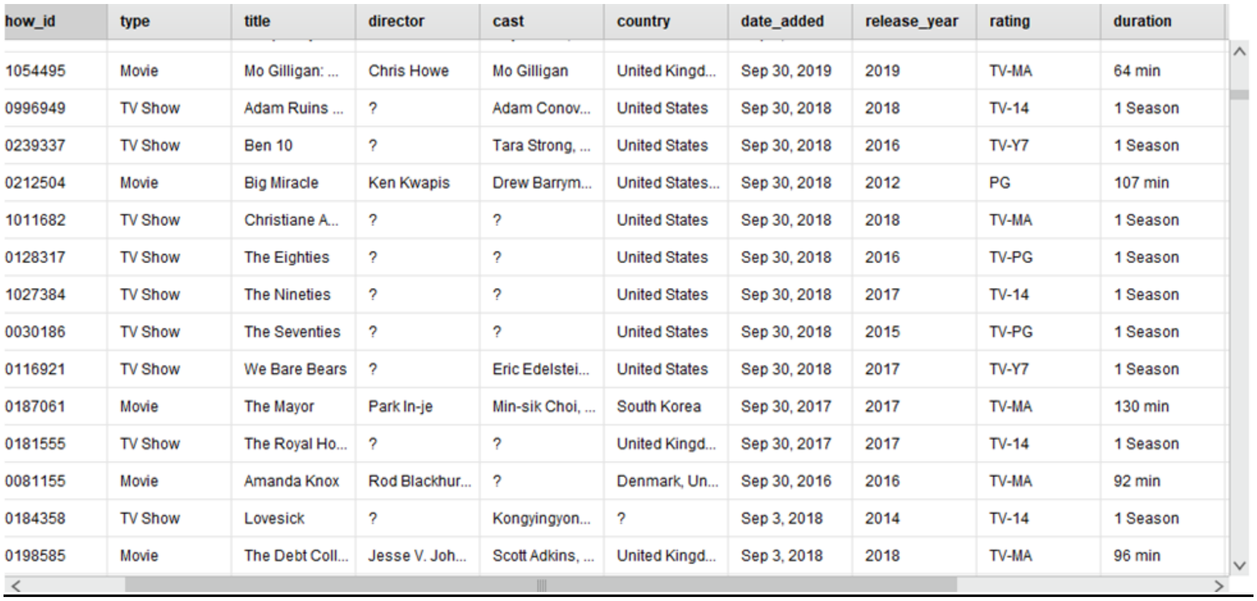
1.0 DATASET & SOURCE



how_id	type	title	director	cast	country	date_added	release_year	rating	duration
1054495	Movie	Mo Gilligan: ...	Chris Howe	Mo Gilligan	United Kingd...	Sep 30, 2019	2019	TV-MA	64 min
0996949	TV Show	Adam Ruins ...	?	Adam Conov...	United States	Sep 30, 2018	2018	TV-14	1 Season
0239337	TV Show	Ben 10	?	Tara Strong, ...	United States	Sep 30, 2018	2016	TV-Y7	1 Season
0212504	Movie	Big Miracle	Ken Kwapis	Drew Barrym...	United States...	Sep 30, 2018	2012	PG	107 min
1011682	TV Show	Christiane A...	?	?	United States	Sep 30, 2018	2018	TV-MA	1 Season
0128317	TV Show	The Eighties	?	?	United States	Sep 30, 2018	2016	TV-PG	1 Season
1027384	TV Show	The Nineties	?	?	United States	Sep 30, 2018	2017	TV-14	1 Season
0030186	TV Show	The Seventies	?	?	United States	Sep 30, 2018	2015	TV-PG	1 Season
0116921	TV Show	We Bare Bears	?	Eric Edelstei...	United States	Sep 30, 2018	2017	TV-Y7	1 Season
0187061	Movie	The Mayor	Park In-je	Min-sik Choi, ...	South Korea	Sep 30, 2017	2017	TV-MA	130 min
0181555	TV Show	The Royal Ho...	?	?	United Kingd...	Sep 30, 2017	2017	TV-14	1 Season
0081155	Movie	Amanda Knox	Rod Blackhur...	?	Denmark, Un...	Sep 30, 2016	2016	TV-MA	92 min
0184358	TV Show	Lovesick	?	Kongyingyon...	?	Sep 3, 2018	2014	TV-14	1 Season
0198585	Movie	The Debt Coll...	Jesse V. Joh...	Scott Adkins, ...	United Kingd...	Sep 3, 2018	2018	TV-MA	96 min

This report analyzes a dataset of Netflix titles, including movies and TV shows. The dataset provides various attributes such as title, director, cast, country, date added, release year, rating, duration, genre, and description. The objective is to explore the dataset to uncover trends and patterns in the types of content available on Netflix.

The dataset comprises multiple attributes that provide detailed information about each title. Here is a brief description of each attribute:

- **show_id**: A unique identifier for each title in the dataset.
- **type**: Indicates whether the title is a "Movie" or a "TV Show".
- **title**: The name of the movie or TV show.
- **director**: The director(s) of the movie or TV show. This field can have multiple directors separated by commas, and it may be blank if the information is not available.
- **cast**: The main actors and actresses in the movie or TV show. This field can also have multiple names separated by commas.
- **country**: The country or countries where the movie or TV show was produced. This field can contain multiple countries separated by commas.
- **date_added**: The date when the movie or TV show was added to Netflix. This is useful for analyzing the trend of content addition over time.

- **release_year**: The year in which the movie or TV show was originally released.
- **rating**: The age rating assigned to the movie or TV show (e.g., TV-MA, TV-14, R, PG-13). This helps in understanding the target audience for each title.
- **duration**: The duration of the movie in minutes or the number of seasons for TV shows.
- **listed_in**: The genres or categories that the movie or TV show belongs to (e.g., Dramas, Comedies, Action & Adventure). This field can contain multiple genres separated by commas.
- **description**: A brief summary or description of the movie or TV show.
- **review :** Sentiment analysis score of the review, indicating whether the review is positive, negative, or neutral.

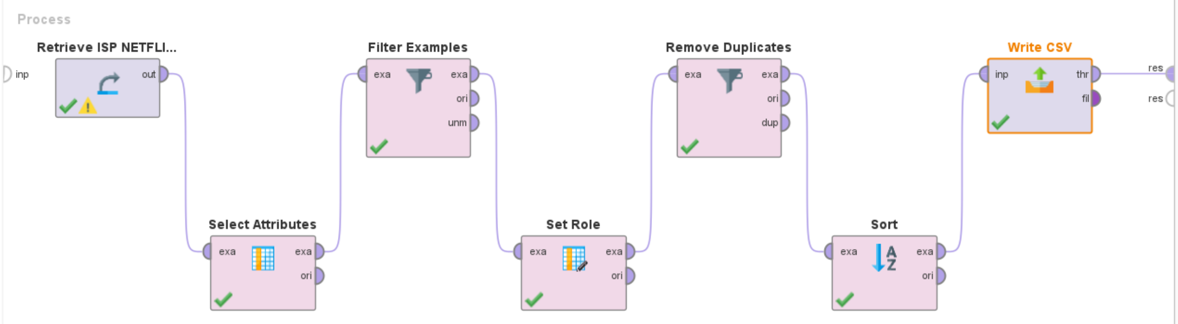
The primary purpose of this dataset is to provide insights into the types of content available on Netflix, the diversity of genres, the origin of the content, and the overall trends in the Netflix catalog. It can be used for various analyses, such as:

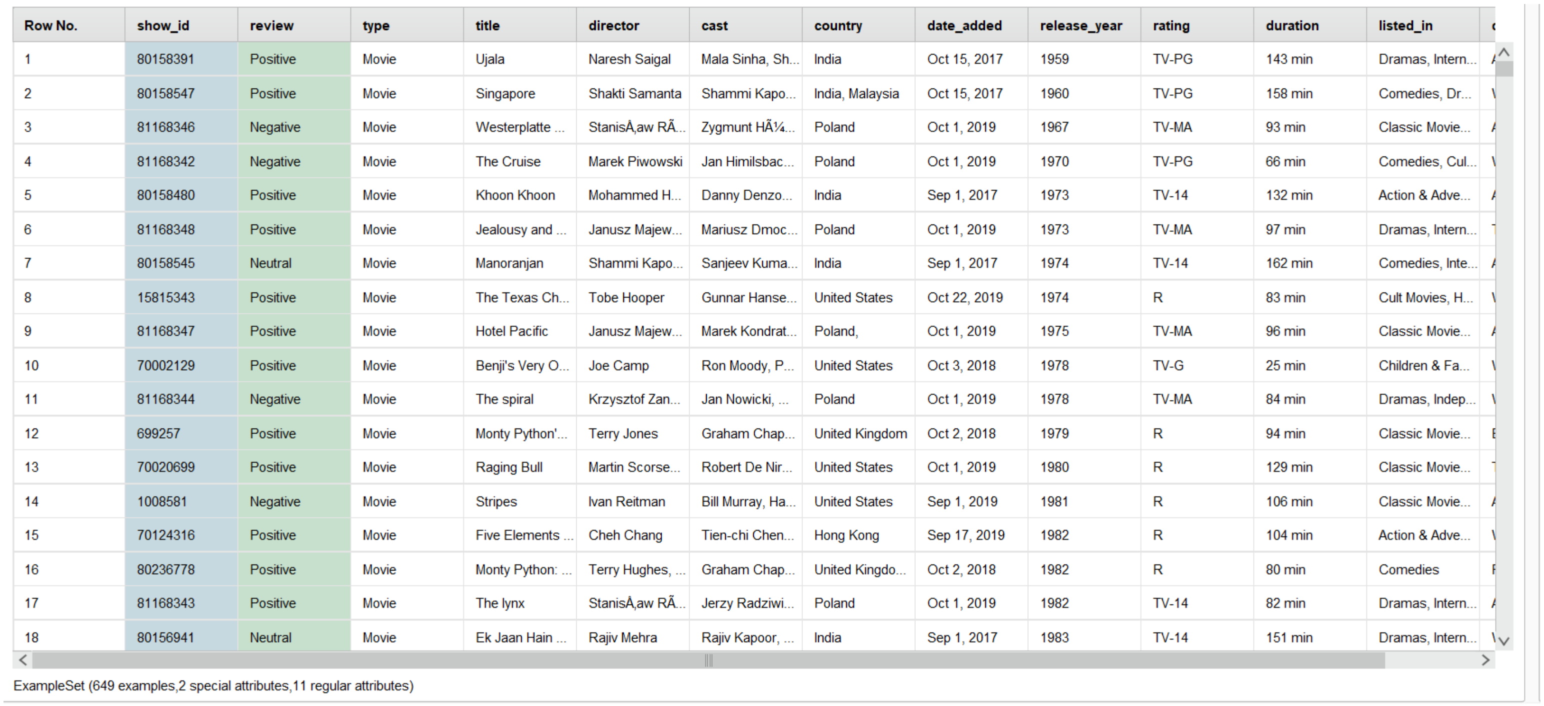
- **Content Analysis**: Understanding the distribution of movies and TV shows, the popularity of different genres, and the representation of various countries.
- **Trend Analysis**: Analyzing how the addition of new content to Netflix has evolved over the years.
- **Audience Analysis**: Examining the ratings to understand the target demographics for Netflix content.
- **Contributor Analysis**: Identifying the most prolific directors and actors in the Netflix catalog.

The dataset is obtained from a trusted source that regularly updates and maintains comprehensive records of Netflix's catalog. While the dataset is reliable and provides a broad overview of the content available on Netflix, it is essential to note that it may not capture real-time changes in the catalog, such as newly added titles or recently removed content. Therefore, the analysis based on this dataset reflects the state of Netflix's catalog at the time the data was collected.

2.0 DESCRIPTIVE ANALYSIS

From the reviews of the Netflix dataset, there are missing values we want to remove from the Netflix data set. Figure below shows the " Retrieve " operator to input the data into the rapidminer. Then we use the " Select Attributes " operator to choose which attributes will appear on the output. "Filter example " operator resulting in clean data which does not include any missings values in the data set. Next, use the " Set Role " operator to set show_id attributes as a unique id or key findings. To reduce data redundancy in the data set this study chooses the " Remove Duplicate " operator to remove any duplicate data. Continue with sorting out the data with the " Sort " operator which sets the movie in ascending order by year. Lastly, " Write CSV " is used to save the clean data into CSV.





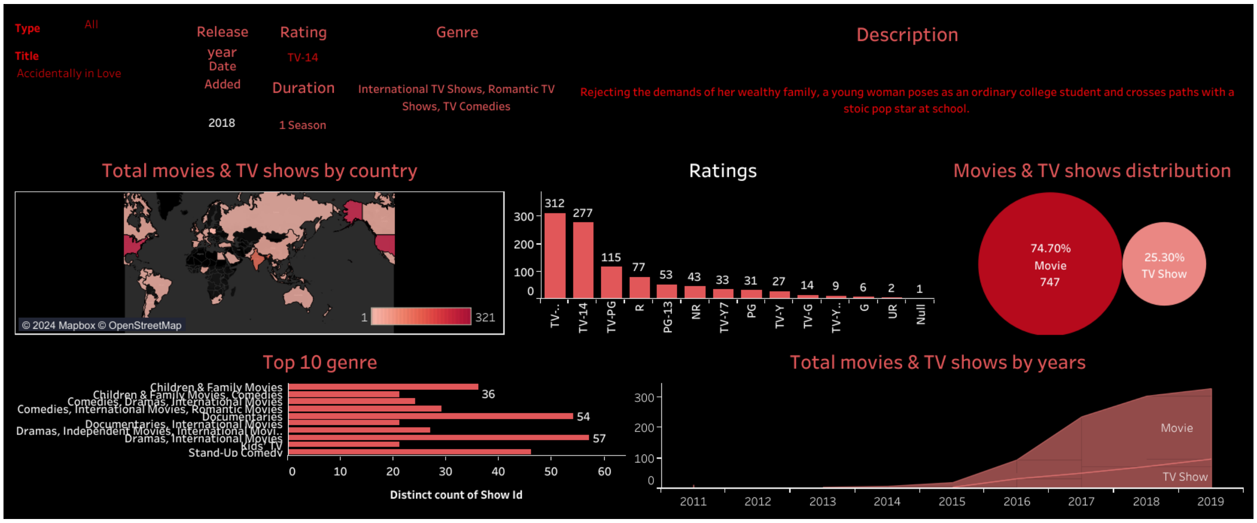
Row No.	show_id	review	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
1	80158391	Positive	Movie	Ujala	Naresh Saigal	Mala Sinha, Sh...	India	Oct 15, 2017	1959	TV-PG	143 min	Dramas, Intern...
2	80158547	Positive	Movie	Singapore	Shakti Samanta	Shammi Kapo...	India, Malaysia	Oct 15, 2017	1960	TV-PG	158 min	Comedies, Dr...
3	81168346	Negative	Movie	Westerplatte ...	StanisÅ‚aw RÃ...	Zygmunt HÃ¼...	Poland	Oct 1, 2019	1967	TV-MA	93 min	Classic Movie...
4	81168342	Negative	Movie	The Cruise	Marek Piwowski	Jan Himilsbac...	Poland	Oct 1, 2019	1970	TV-PG	66 min	Comedies, Cul...
5	80158480	Positive	Movie	Khoon Khoon	Mohammed H...	Danny Denzo...	India	Sep 1, 2017	1973	TV-14	132 min	Action & Adve...
6	81168348	Positive	Movie	Jealousy and ...	Janusz Majew...	Mariusz Dmoc...	Poland	Oct 1, 2019	1973	TV-MA	97 min	Dramas, Intern...
7	80158545	Neutral	Movie	Manoranjan	Shammi Kapo...	Sanjeev Kuma...	India	Sep 1, 2017	1974	TV-14	162 min	Comedies, Inte...
8	15815343	Positive	Movie	The Texas Ch...	Tobe Hooper	Gunnar Hanse...	United States	Oct 22, 2019	1974	R	83 min	Cult Movies, H...
9	81168347	Positive	Movie	Hotel Pacific	Janusz Majew...	Marek Kondrat...	Poland,	Oct 1, 2019	1975	TV-MA	96 min	Classic Movie...
10	70002129	Positive	Movie	Benji's Very O...	Joe Camp	Ron Moody, P...	United States	Oct 3, 2018	1978	TV-G	25 min	Children & Fa...
11	81168344	Negative	Movie	The spiral	Krzysztof Zan...	Jan Nowicki, ...	Poland	Oct 1, 2019	1978	TV-MA	84 min	Dramas, Indep...
12	699257	Positive	Movie	Monty Python'...	Terry Jones	Graham Chap...	United Kingdom	Oct 2, 2018	1979	R	94 min	Classic Movie...
13	70020699	Positive	Movie	Raging Bull	Martin Scorse...	Robert De Nir...	United States	Oct 1, 2019	1980	R	129 min	Classic Movie...
14	1008581	Negative	Movie	Stripes	Ivan Reitman	Bill Murray, Ha...	United States	Sep 1, 2019	1981	R	106 min	Classic Movie...
15	70124316	Positive	Movie	Five Elements ...	Cheh Chang	Tien-chi Chen...	Hong Kong	Sep 17, 2019	1982	R	104 min	Action & Adve...
16	80236778	Positive	Movie	Monty Python: ...	Terry Hughes, ...	Graham Chap...	United Kingdo...	Oct 2, 2018	1982	R	80 min	Comedies
17	81168343	Positive	Movie	The lynx	StanisÅ‚aw RÃ...	Jerzy Radziwi...	Poland	Oct 1, 2019	1982	TV-14	82 min	Dramas, Intern...
18	80156941	Neutral	Movie	Ek Jaan Hain ...	Rajiv Mehra	Rajiv Kapoor, ...	India	Sep 1, 2017	1983	TV-14	151 min	Dramas, Intern...

ExampleSet (649 examples,2 special attributes,11 regular attributes)

3.0 DIAGNOSTIC ANALYSIS

Diagnostic analytics are used for discovery or to determine why something happened. Data analysts frequently utilize Tableau to develop and share interactive business intelligence dashboards online in real-time, accessible through a web browser by others within their organization. All ratings of the Netflix movie dataset were integrated with Tableau for analysis.

DASHBOARD



Type All

Title Accidentally in Love

Release year Date Added 2018

Rating TV-14

Duration 1 Season

Genre International TV Shows, Romantic TV Shows, TV Comedies

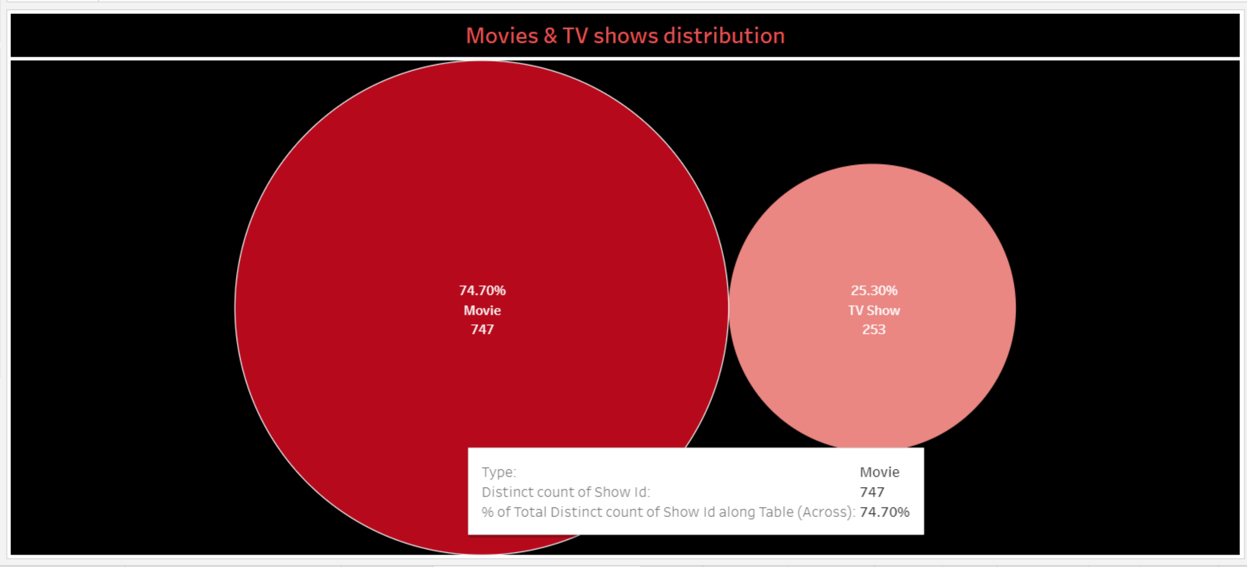
Description Rejecting the demands of her wealthy family, a young woman poses as an ordinary college student and crosses paths with a stoic pop star at school.

Total movies & TV shows by country
© 2024 Mapbox © OpenStreetMap
1
321
Ratings
300
200
100
0
312
277
115
77
53
43
33
31
27
14
9
6
2
1
TV-..
TV-14
TV-PG
R
PG-13
NR
TV-Y7
PG
TV-Y
TV-G
TV-Y..
G
UR
Null
Movies & TV shows distribution
74.70%
Movie
747
25.30%
TV Show
Top 10 genre
Children & Family Movies
Children & Family Movies, Comedies
Comedies, Dramas, International Movies
Comedies, International Movies, Romantic Movies
Documentaries
Documentaries, International Movies
Dramas, Independent Movies, International Movi..
Dramas, International Movies
Kids' TV
Stand-Up Comedy
36
54
57
0
10
20
30
40
50
60
Distinct count of Show Id
Total movies & TV shows by years
300
200
100
0
Movie
TV Show
2011
2012
2013
2014
2015
2016
2017
2018
2019

This dashboard contains 5 sheets of data visualization and describes all the data in Netflix dataset. This visualization effectively provides insights into the distribution, growth, and types of content available on Netflix, helping to understand the platf orm's content strategy and viewer preferences. From this visual, analysis can be understood clearly and easily.

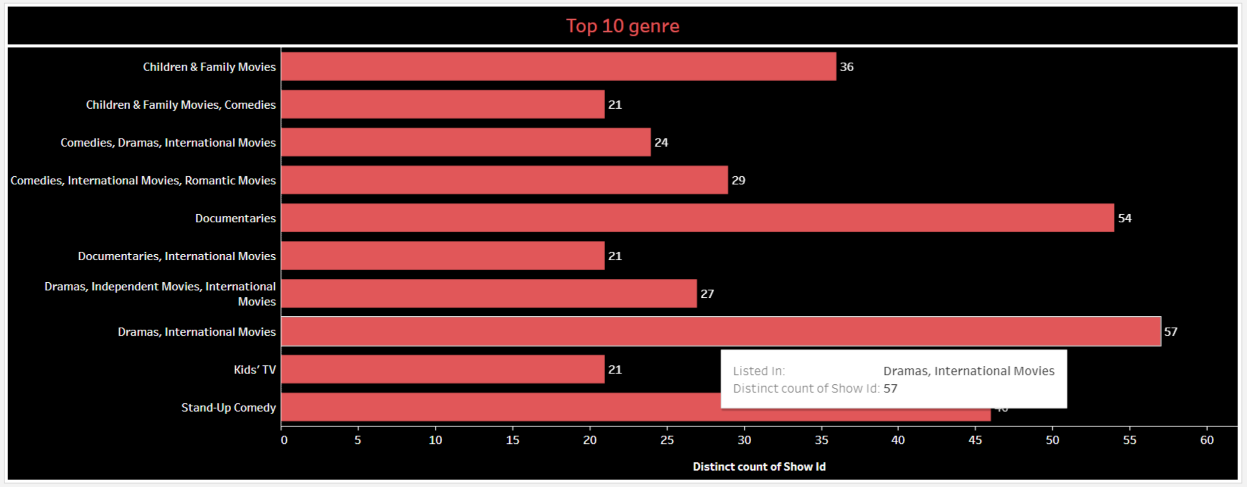
STORY TELLING

Story 1



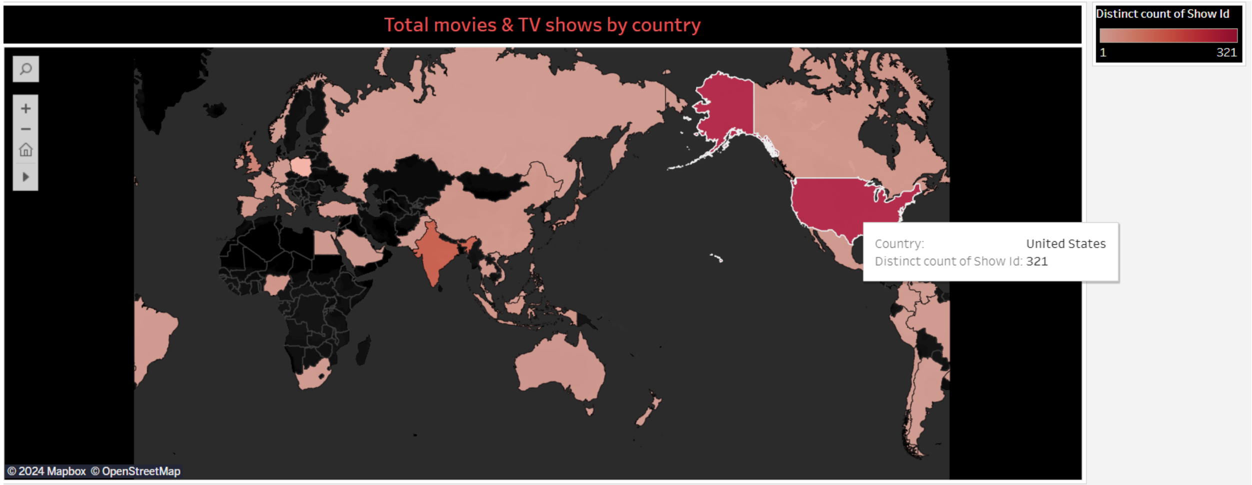
- ➢ There is a significantly higher number of movies with 747 titles compared to TV shows.

Story 2



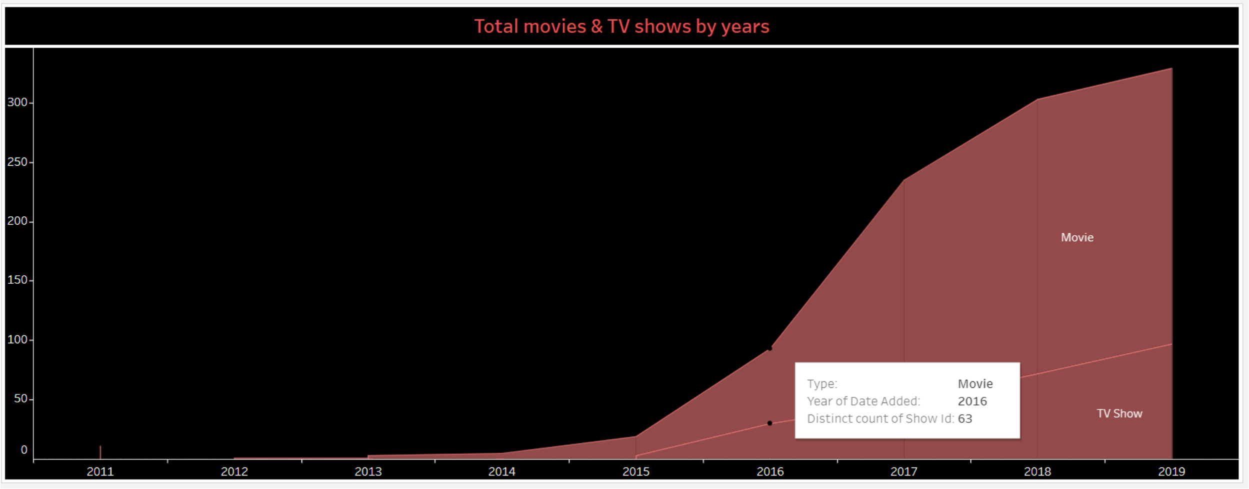
- ➢ International Movies and Dramas are among the top genres with 57 titles, highlighting the diverse content available on the platform.

Story 3


Total movies & TV shows by country
Distinct count of Show Id
1
321
Country: United States
Distinct count of Show Id: 321
© 2024 Mapbox © OpenStreetMap

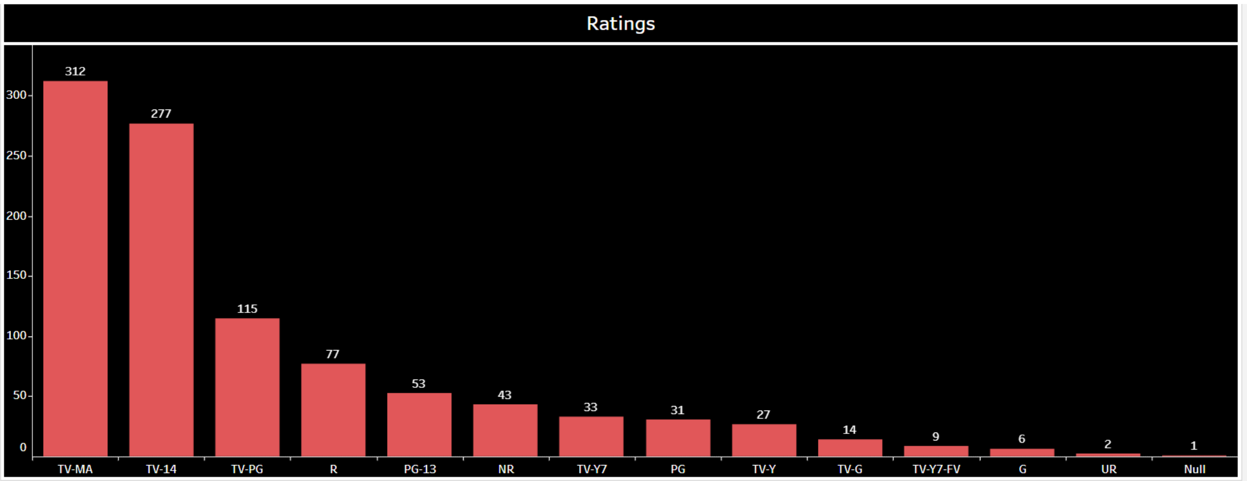
- ➢ The United States has the highest concentration of available content with 321 titles, followed by other regions with lighter shades on the map.

Story 4



- ➢ The addition of content has been increasing over the years, from 2011 to 2015, content added each year increased steadily. By 2016 saw the beginning of a sharp rise in content addition. By 2019, there was a significant peak in new content.

Story 5

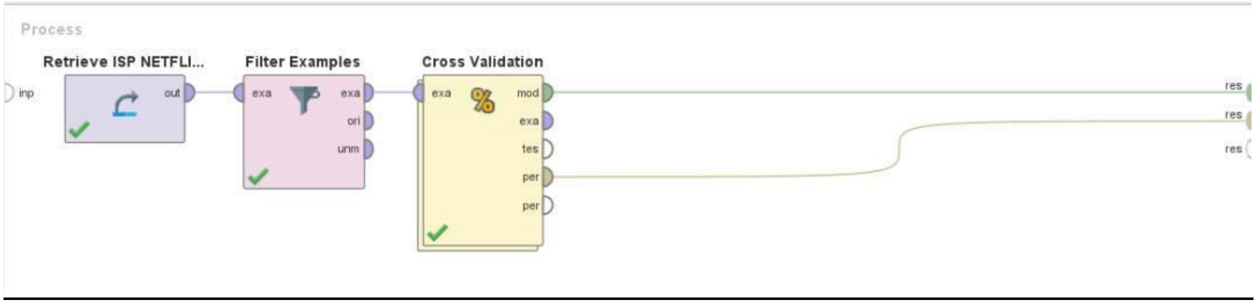


- The most common ratings are TV-MA **(Mature Audiences Only)** with 312 titles and TV-14 **(Parents Strongly Cautioned)** with 277 titles, indicating a preference for content suitable for mature audiences.

4.0 PREDICTIVE ANALYSIS

The use of statistics and modeling approaches for predicting future results and performance is known as predictive analytics. With predictive analytics, data patterns in the past and present are examined to see if they are likely to recur. The model has been used to determine the data's accuracy and to boost confidence in the validity and reliability of the results for the Netflix dataset. We have chosen a Decision Tree model for this project. Based on the probability calculation value of the chosen models, the classification procedure assigns the labels positive, negative, and neutral.

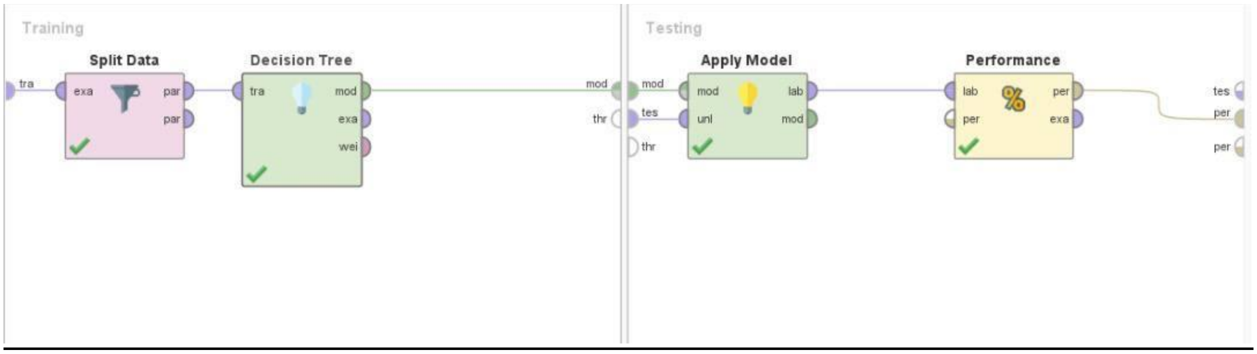
These diagrams below show the workflow in RapidMiner. Firstly, we need to "Retrieve" the Netflix dataset by dragging it into the process to copy what is received on input to all its output ports. Then, we use the "Filter Examples" operator to remove outliers or select specific subsets of data relevant to the analysis. Basically , we will filter some data that is missing to make the system become easier to read. Next, we are going to do the "Cross Validation" operator to apply the Decision Tree Model. This technique is used to evaluate the effectiveness of a statistical method. It splits the dataset into two segments which are training and testing sets. All these operators must be connected to the output ports.



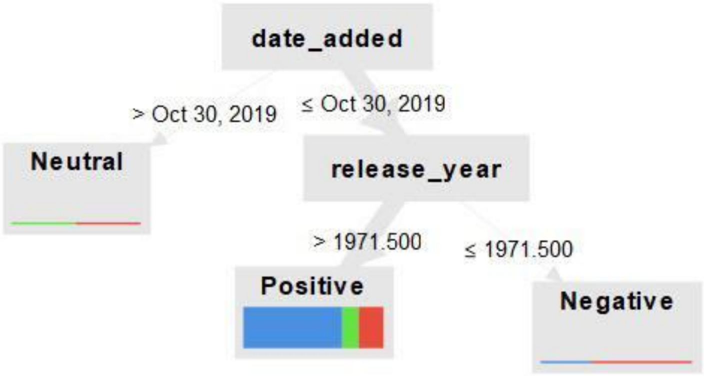
For the training set, there are two operators we use in the process named "Split Data" operator and "Decision tree" operator. "Split data" operator takes an ExampleSet as its input and delivers the subsets of that ExampleSet through its output ports. The number of subsets (or partitions) and the relative size of each partition are specified through the partitions parameter. We use the ratio 0.7 : 0.3 for our project to analyze our data. The sum of the ratio should be 1. Then we use the "Decision Tree" operators to apply the Decision Tree Model.

Other than that, the testing set also has two operators which are “Apply model” operator and “Performance” operator. “Apply Model” operator applies a model on an ExampleSet. This node applies the trained model to the test data. Usually, the goal is to get a prediction on unseen data or to transform data by applying a preprocessing model. The model’s predictions are then compared to the actual outcomes. The “Performance” operator is to evaluate the performance of the model using metrics that could include accuracy, precisions, recall and other. This operator is used for statistical performance evaluation of classification tasks. This operator also delivers a list of performance criteria values of the classification task. The “Performance” (Classification) operator is used with classification tasks only. On the other hand, the “Performance” operator automatically determines the learning task type and calculates the most common criteria for that type. Classification is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the train on a particular day will be 'on time', 'late' or 'very late'.

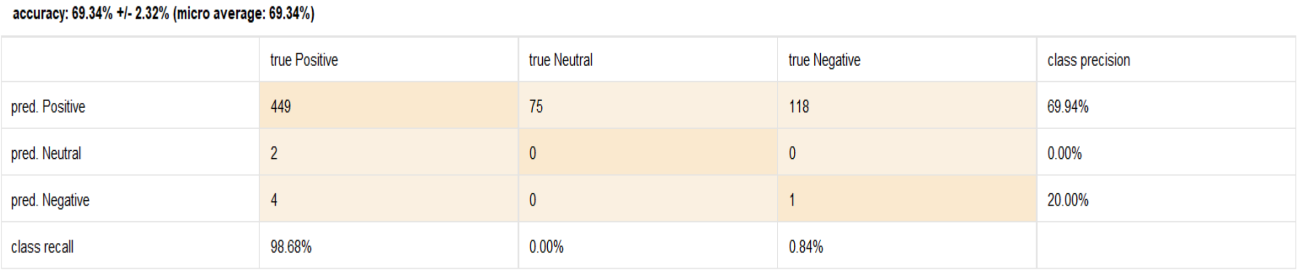
Each of these steps is crucial for developing a predictive model that is both accurate and generalizable to new data. RapidMiner provides a visual and interactive environment to streamline these processes, making it accessible for users with varying levels of programming expertise.



Decision trees are used to solve classification problems and categorize objects depending on their learning features. They can also be used for regression problems or as a method to predict continuous outcomes from unforeseen data.



5.0 CONCLUSION



accuracy: 69.34% +/- 2.32% (micro average: 69.34%)

	true Positive	true Neutral	true Negative	class precision
pred. Positive	449	75	118	69.94%
pred. Neutral	2	0	0	0.00%
pred. Negative	4	0	1	20.00%
class recall	98.68%	0.00%	0.84%	

The model displays the accuracy of Netflix reviews by using the Decision Tree Model. The accuracy includes all the reviews that are positive, negative and neutral. The total of true positives is 449, the total of true negatives are 1 and the total of the true neutral reviews are 0 while the total observations is 649. Therefore, the accuracy is 69.34% .The class Recall for Positive is 98.68%,the model is very effective at identifying positive cases, missing very few. While , the class precision for Positive is 69.94%.Of the cases predicted as positive, about 70% were actually positive. Both the class neutral and class precision was 0.00%. The class recall for Negative and class precision for Negative is 0.84% and 20.00%.